CMPE 239: Web and Data Mining
Project Report

Classification of Restaurants from Customer Reviews

Submitted To Professor David C. Anastasiu

Submitted By Group 15

Raghavendra Guru 010711974
Navit Gaur 010702562
Shubham Vadhera 010152545

Section 1
Introduction

Motivation

Review Services like Yelp, Google Reviews etc. provide customers and businesses a way to interact with one another. Reviews and Ratings are useful source of information but significant problem exists in extracting relevant information and predicting the future through analysis and correlation of the existing data. Yelp is currently the leading review service provider in the market. Each day thousands of restaurants and businesses are reviewed by the customers. Currently there is no efficient way to categorize the restaurants based on customer reviews.

Objective

In our project we try to predict category of the restaurant based on customer reviews. Our system will help other customers make decision on the restaurant to select based on the type of food they are interested in. The best way to categorise a restaurant is using customer reviews. The suggestions based on this approach are closer to the likings of the users. Our intention is to explore this approach and predict food category based on reviews that matches the actual food category.

Literature Survey

Consumer review websites such as Yelp, TripAdvisor etc. have become increasingly popular over the past decade, and now exist for nearly every product and service. Yelp alone contains more than 2.6 million reviews of restaurants and has a market capitalization of roughly two billion dollars. Moreover, there is mounting evidence that these reviews have a direct influence on product sales.

The dataset used for our project is the one available as a part of the Yelp dataset challenge (http://www.yelp.com/dataset_challenge). Most of the predictive tasks previously performed on this dataset have category predictions primarily based on business attribute [1]. However research has been carried out, not just in the general area of text mining and sentiment analysis, but in text mining for predictive tasks in review and rating systems. The impact of text derived information has been previously studied at the sentence level with the help of the topic information on various datasets [1].

Various methods have been adopted in the past, including regression[2], bag of opinions method [3] etc. In food reviews, it has been observed that Naive Bayes' had a slightly better accuracy than the SVM method. However, this was in combination with other features of the

dataset. Hence, the results differ from the ones chosen here. The Yelp dataset has been extensively studied as well. Attempts have been made to gauge information from the review text by predicting what the user felt about various aspects of the business, such as service, quality of food and ambiance. [4] [5] If the user experience can be divided into various aspects, then a function of these can be used to predict the overall food category.

Section 2
System Design & Implementation details

1. Algorithms considered/selected:

We have used KNN based approach for our classification model. We have also used K- fold cross validation (K=10) to validate our results. This algorithm helped us eliminate overfitting from our project. We have used PCA for dimensionality reduction. It is a simple, fast and efficient algorithm to reduce the size of our data set.

2. Technologies & Tools used:

We used Python as our programming language since it has many pre written libraries which can be used to perform KNN classification as well as 10 fold cross validation. To run our scripts we used jupyter notebook and also the command line occasionally. We have also used PyCharm IDE to develop the scripts. We have also used python libraries like Numpy, Scipy, NLTK, Pandas, Scikit-learn, Subprocess, Json etc.

3. System design:

Our system consists of multiple approach. Our initial Dataset is the raw data downloaded from Yelp. Then we filter out the data according to our requirements and store just the necessary dataset. Then we create a term frequency vector for our data and process the vector on our developed classification model. To further reduce the size of the data we perform dimensionality reduction on the vector. Then again we use the same model for classification. At the end we perform 10 fold cross validation on various dataset we had. The performance of all the approach was compared.

Our System has following components:

- Data: Our data comes from Yelp. It contains 2.7M reviews and about 86K businesses. We then pre process the data. The result is two different datasets. One of them contains all the review text without common english words, stop words, punctuations etc. The second dataset contains all the review texts related to food words.

Then we generate term frequency vector for both the approach. Dimensionality reduction is performed on both the vectors to generate two more vectors of reduced size.

- Classification Model: We then perform classification for the four datasets generated. Our classification model uses KNN based approach. We classify terms nearest to each other into a single label.
- 10 fold cross validation: To validate our results we perform 10 fold cross validation. Validation is necessary since we need to verify if our model generates the desired results or not.
- Performance Analysis: We compare the results and accuracy of all the four approach. We assess the performance and accuracy for our vectors.

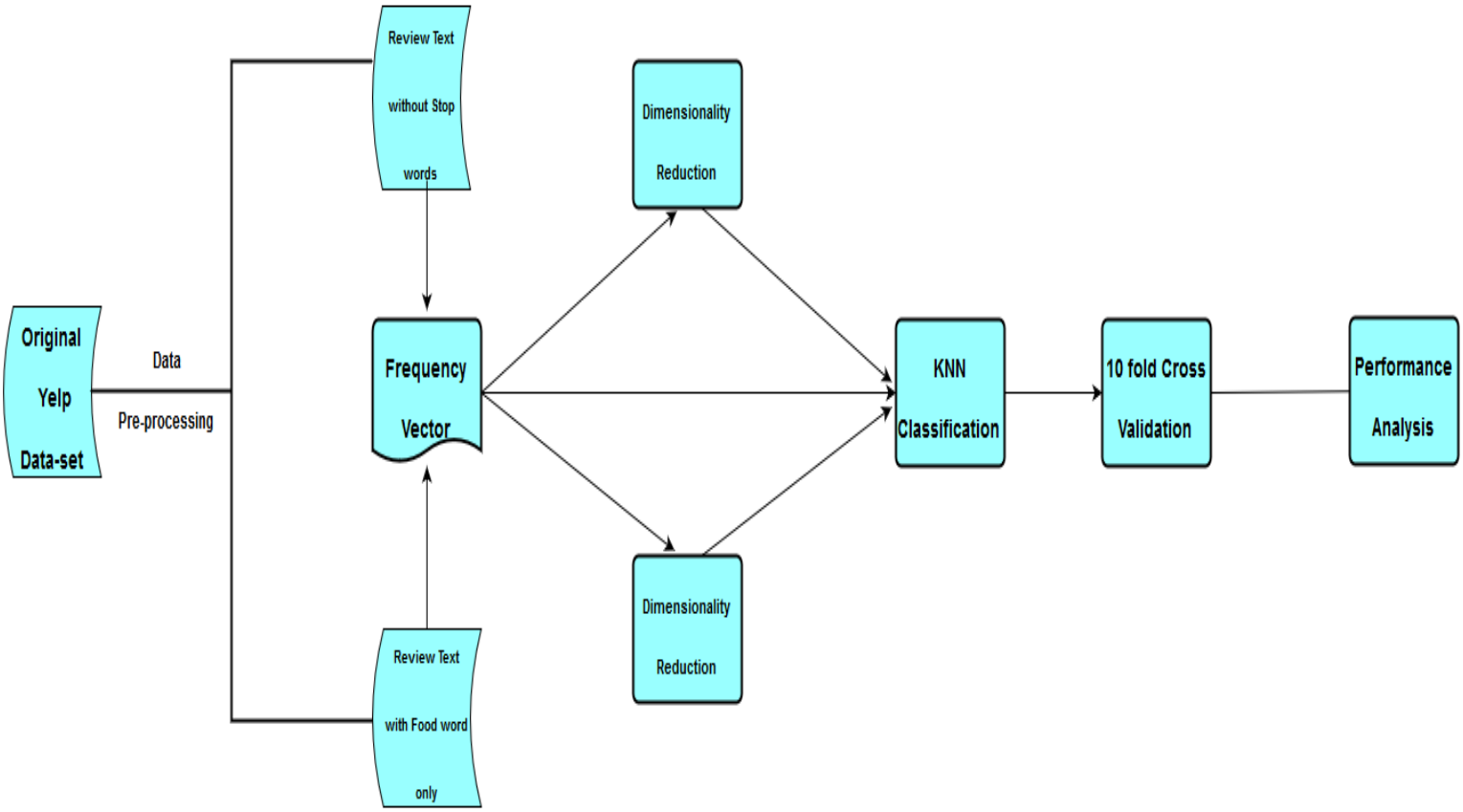

Figure1: Data/Operation Flow for our system

Section 3
Experiments / Proof of concept evaluation

Dataset used:

Yelp Academic Dataset

- 2.7M reviews and 649K tips by 687K users for 86K businesses.
- Downloaded from Yelp Dataset Challenge : https://www.yelp.com/dataset_challenge
- All the data is in JSON format.
- We considered all the 2.7M Reviews, 86K Businesses and the categories related to them.
- Since the quantity of data is huge. We filtered out the data which was required.

Pre-processing:

- Web Scraper: We developed a web scraper which considers the names of all the food items based on their ID from the website http://allrecipes.com/recipe/ and automatically stores all the food items into a text file.
- Then in the second step we removed the repeated food items and generated a file containing unique food items.
- For Approach 1: We removed all the stop words, punctuations and commonly used english language words from the review text.
- For Approach 2: We considered only the common food words generated in Approach 1 and removed any other additional words from the review text.

Methodology followed:

1. In our project we decided four different Approach and prepared the dataset for each :

Approach 1:

- Prepare term frequency vector for dataset containing all the reviews excluding stop words, punctuations, commonly used english words.

Approach 2:

- Prepare term frequency vector for dataset containing all the reviews but we include only food related terms.

Approach 3:

- Perform dimensionality reduction on dataset obtained in Approach 1.

Approach 4:

- Perform dimensionality reduction on dataset obtained in Approach 2.

2. Build the model using training data set.
3. Test the model using the test data:
    a. Compute distance to other training records.

b. Identify k nearest neighbors in training set.

c. Use class labels of nearest neighbors to determine the class label.

4. The Step 2 and 3 performed on all the 4 data sets generated in step 1.
5. K-fold was used to cross validate all the models.
6. Performance of all the approaches were compared against each other.

Analysis of result

We have compared the results we got after using our classification model on the complete dataset for all the approach we followed. We have also compared the result we got for randomly selected data based on our various approach.

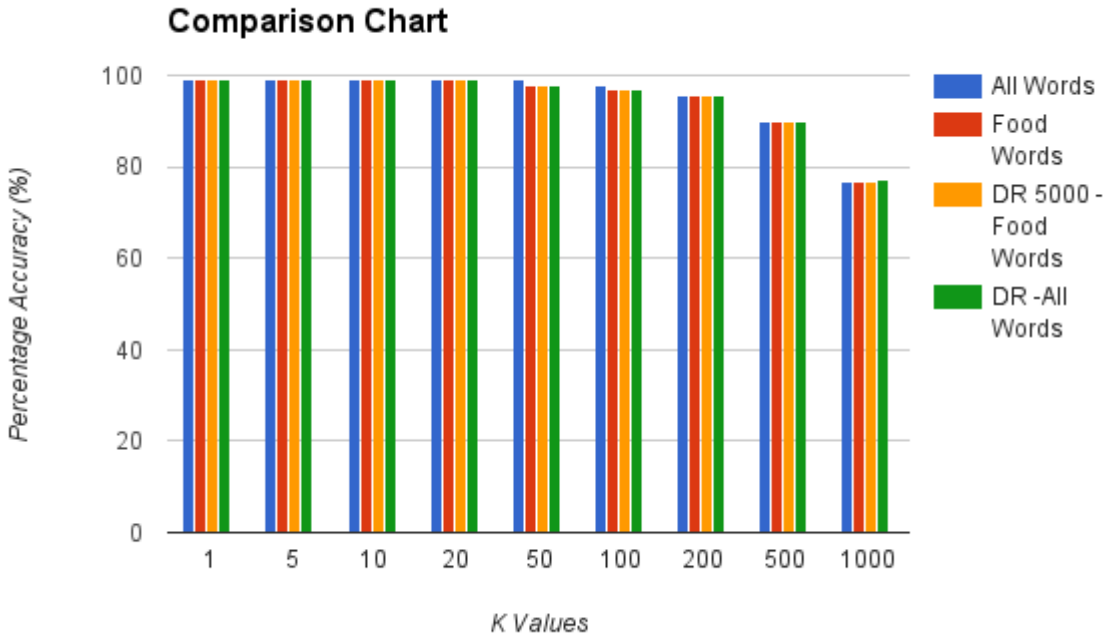

Figure2: Comparison of accuracy for various approach

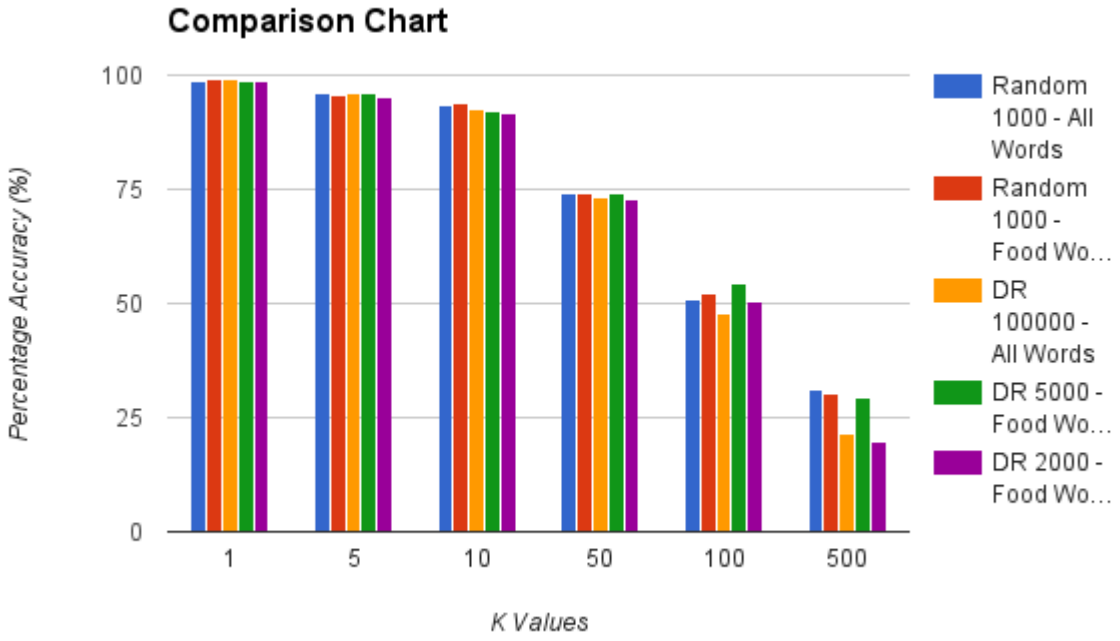

Figure3: Comparison of accuracy for random data for various approach

Section 4
Discussion & Conclusions

Decisions made:

- Since the review data was too big we decide to remove common english words and stopwords to reduce the dataset.
- We tried performing classification with both naive bayes and KNN method but since KNN is better performance wise and produces better results we decided to use KNN method.
- We considered only the top few class labels that we extracted ignoring less used food categories since it would not have made significant difference to our results.

Difficulties faced:

- The complete dataset itself was huge and it took a lot of time to process the data.
- Extracting all the food items from the website automatically was challenging. We checked for close to 300000 food items based on their ID. Although some of them were empty but still checking 300000 items and writing them to a file is time consuming task. We also had to tackle network outages, system failures while

running the script. For our project it took almost 3 days of continuously running the scrapper parallelly on 3 systems.

- Running the complete script was time as well as memory consuming. To tackle this issue, we divided the tasks into many small scripts, and we would save files after each step.

Things that worked:

- Creating a script to scrape food words from a food website worked well since we got sufficient data to extract food related words from our dataset.
- 10 fold cross validation helped us validate our results and compare the prediction accuracy of our four approach.

Things that didn't work well :

- Word2Vec - pre-trained on Google News
- Sklearn - feature_extraction.text.HashingVectorizer
- Sklearn - naive_bayes import MultinomialNB

Difficulties faced:

- Data set which has only food words give better result (0.2%) than that of the complete data set.
- Reducing the dimensions of the data didn't improve the accuracy of the Classification but it ran much faster.
- Keeping K=1,5,10 gives the better result compared to k=100.

Section 5
Project Plan / Task Distribution

Tasks Assigned:

Raghavendra Guru:

- Creating a frequency vector for the reviews obtained in Step 1 and Step 3.
- Using these 4 vectors to perform classification using KNN approach and predicting the appropriate class label for the text.
- Dimensionality reduction.
- Build classification model and Validate the model.

Shubham Vadhera:

- Figuring out how to identify food words
- Decide class labels, categories.
- 10-fold cross validation
- Results and performance comparison approach

Navit Gaur:

- Extracting Reviews and removing stop words, punctuations and commonly used english words and storing them into a separate file.
- Extracting all the unique categories and their count from the businesses provided in Yelp dataset. Removing less frequent food categories and keeping just the relevant categories as Class Labels.
- Performing Dimensionality Reduction on both the frequency vectors.

Tasks Performed:

Raghavendra Guru:

- Combine the review and category json file and generate a single file then generate BOW (Bag of words).
- Perform Dimensionality reduction on term frequency vector.
- Using these 4 vectors to perform classification using KNN approach and predicting the appropriate class label for the text.
- Performing 10 - fold cross validation to validate the prediction accuracy of the model developed.
- Analyze the results of our implementation and compare different approaches.

Shubham Vadhera:

- Worked on food scraper, cleaning source files for food words.
- Script that decides one class label amongst multiple class labels on the basis of popularity in reviews
- Worked on script that performs 10 - fold cross validation to validate the prediction accuracy of the model developed.
- Worked on random selection dimensionality reduction.
- Comparing graphically the results and performance for all the four approach.

Navit Gaur:

- Extracted Reviews and removed stop words, punctuations and commonly used 5000 english words and stored them into a separate file.
- Performed data pre-processing to generate the required datasets to be used by our model and also worked on the project report.
- Extracted all the unique categories and their count from the businesses provided in Yelp dataset.
- Removed less frequent food categories and kept just the relevant categories as Class Labels.

References:

[1] Technical Report: https://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/031.pdf
[2] Ganu, Gayatree, Noemie Elhadad, and Amelie Marian. "Beyond the Stars: Improving Rating Predictions using Review Text Content." WebDB. Vol. 9. 2009.
[3] Qu, Lizhen, Georgiana Ifrim, and Gerhard Weikum. "The bag-of-opinions method for review rating prediction from sparse text patterns." Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.
[4] Technical Report: http://www.ics.uci.edu/~vpsaini/ les/technical_report.pdf
[5] Yelp Challenge Presentation: http://www.ics.uci.edu/~vpsaini/
[6]Scikit-Learn:http://scikitlearn.org/stable/tutorial/text_analytics/working_with_text_data.html